Unsupervised Visual Sense Disambiguation for Verbs using Multimodal Embeddings

Spandana Gella, Mirella Lapata and Frank Keller
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
S.Gella@sms.ed.ac.uk, mlap@inf.ed.ac.uk, keller@inf.ed.ac.uk

Abstract

We introduce a new task, visual sense disambiguation for verbs: given an image and a verb, assign the correct sense of the verb, i.e., the one that describes the action depicted in the image. Just as textual word sense disambiguation is useful for a wide range of NLP tasks, visual sense disambiguation can be useful for multimodal tasks such as image retrieval, image description, and text illustration. We introduce VerSe, a new dataset that augments existing multimodal datasets (COCO and TUHOI) with sense labels. We propose an unsupervised algorithm based on Lesk which performs visual sense disambiguation using textual, visual, or multimodal embeddings. We find that textual embeddings perform well when gold-standard textual annotations (object labels and image descriptions) are available, while multimodal embeddings perform well on unannotated images. We also verify our findings by using the textual and multimodal embeddings as features in a supervised setting and analyse the performance of visual sense disambiguation task. VerSe is made publicly available and can be downloaded at: `https://github.com/spandanagella/verse`.

1 Introduction

Word sense disambiguation (WSD) is a widely studied task in natural language processing: given a word and its context, assign the correct sense of the word based on a pre-defined sense inventory (Kilgarrif, 1998). WSD is useful for a range of NLP tasks, including information retrieval, information extraction, machine translation, content analysis, and lexicography (see Navigli (2009) for an overview).

Figure 1: Visual sense ambiguity: three of the senses of the verb *play*.

Standard WSD disambiguates words based on their *textual context*; however, in a multimodal setting (e.g., newspaper articles with photographs), *visual context* is also available and can be used for disambiguation. Based on this observation, we introduce a new task, *visual sense disambiguation* (VSD) for verbs: given an image and a verb, assign the correct sense of the verb, i.e., the one depicted in the image. While VSD approaches for nouns exist, VSD for verbs is a novel, more challenging task, and related in interesting ways to *action recognition* in computer vision. As an example consider the verb *play*, which can have the senses *participate in sport*, *play on an instrument*, and *be engaged in playful activity*, depending on its visual context, see Figure 1.

We expect visual sense disambiguation to be useful for multimodal tasks such as image retrieval. As an example consider the output of Google Image Search for the query *sit*: it recognizes that the verb has multiple senses and tries to cluster relevant images. However, the result does not capture the polysemy of the verb well, and would clearly benefit from VSD (see Figure 2).

Visual sense disambiguation has previously been attempted for nouns (e.g., *apple* can mean *fruit* or *computer*), which is a substantially easier task that can be solved with the help of an object detector

Proceedings of NAACL-HLT 2016, pages 182–192,
San Diego, California, June 12-17, 2016.

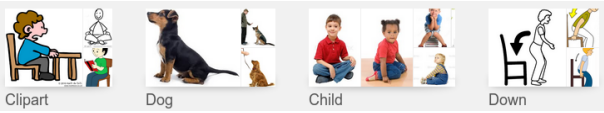

Figure 2: Google Image Search trying to disambiguate *sit*. All clusters pertain to the *sit down* sense, other senses (*baby sit*, *convene*) are not included.

(Barnard et al., 2003; Loeff et al., 2006; Saenko and Darrell, 2008; Chen et al., 2015). VSD for nouns is helped by resources such as ImageNet (Deng et al., 2009), a large image database containing 1.4 million images for 21,841 noun synsets and organized according to the WordNet hierarchy. However, we are not aware of any previous work on VSD for verbs, and no ImageNet for verbs exists. Not only image retrieval would benefit from VSD for verbs, but also other multimodal tasks that have recently received a lot of interest, such as automatic image description and visual question answering (Karpathy and Li, 2015; Fang et al., 2015; Antol et al., 2015).

In this work, we explore the new task of visual sense disambiguation for verbs: given an image and a verb, assign the correct sense of the verb, i.e., the one that describes the action depicted in the image. We present VerSe, a new dataset that augments existing multimodal datasets (COCO and TUHOI) with sense labels. VerSe contains 3518 images, each annotated with one of 90 verbs, and the OntoNotes sense realized in the image. We propose an algorithm based on the Lesk WSD algorithm in order to perform unsupervised visual sense disambiguation on our dataset. We focus in particular on how to best represent word senses for visual disambiguation, and explore the use of textual, visual, and multimodal embeddings. Textual embeddings for a given image can be constructed over object labels or image descriptions, which are available as gold-standard in the COCO and TUHOI datasets, or can be computed automatically using object detectors and image description models.

Our results show that textual embeddings perform best when gold-standard textual annotations are available, while multimodal embeddings perform best when automatically generated object labels are used. Interestingly, we find that automatically generated image descriptions result in inferior performance.

Dataset	Verbs	Acts	Images	Sen	Des
PPMI (Yao and Fei-Fei, 2010)	2	24	4800	N	N
Stanford 40 Actions (Yao et al., 2011)	33	40	9532	N	N
PASCAL 2012 (Everingham et al., 2015)	9	11	4588	N	N
89 Actions (Le et al., 2013)	36	89	2038	N	N
TUHOI (Le et al., 2014)	–	2974	10805	N	N
COCO-a (Ronchi and Perona, 2015)	140	162	10000	N	Y
HICO (Chao et al., 2015)	111	600	47774	Y	N
VerSe (our dataset)	90	163	3518	Y	Y

Table 1: Comparison of VerSe with existing action recognition datasets. Acts (actions) are verb-object pairs; Sen indicates whether sense ambiguity is explicitly handled; Des indicates whether image descriptions are included.

2 Related Work

There is an extensive literature on word sense disambiguation for nouns, verbs, adjectives and adverbs. Most of these approaches rely on lexical databases or sense inventories such as WordNet (Miller et al., 1990) or OntoNotes (Hovy et al., 2006). Unsupervised WSD approaches often rely on distributional representations, computed over the target word and its context (Lin, 1997; McCarthy et al., 2004; Brody and Lapata, 2008). Most supervised approaches use sense annotated corpora to extract linguistic features of the target word (context words, POS tags, collocation features), which are then fed into a classifier to disambiguate test data (Zhong and Ng, 2010). Recently, features based on sense-specific semantic vectors learned using large corpora and a sense inventory such as WordNet have been shown to achieve state-of-the-art results for supervised WSD (Rothe and Schutze, 2015; Jauhar et al., 2015).

As mentioned in the introduction, all existing work on visual sense disambiguation has used nouns, starting with Barnard et al. (2003). Sense discrimination for web images was introduced by Loeff et al. (2006), who used spectral clustering over multimodal features from the images and web text. Saenko and Darrell (2008) used sense definitions in a dictionary to learn a latent LDA space overs senses, which they then used to construct sense-specific classifiers by exploiting the text surrounding an image.

2.1 Related Datasets

Most of the datasets relevant for verb sense disambiguation were created by the computer vision community for the task of human action recognition (see

Table 1 for an overview). These datasets are annotated with a limited number of actions, where an action is conceptualized as verb-object pair: *ride horse*, *ride bicycle*, *play tennis*, *play guitar*, etc. Verb sense ambiguity is ignored in almost all action recognition datasets, which misses important generalizations: for instance, the actions *ride horse* and *ride bicycle* represent the same sense of *ride* and thus share visual, textual, and conceptual features, while this is not the case for *play tennis* and *play guitar*. This is the issue we address by creating a dataset with explicit sense labels.

VerSe is built on top of two existing datasets, TUHOI and COCO. The Trento Universal Human-Object Interaction (TUHOI) dataset contains 10,805 images covering 2974 actions. Action (human-object interaction) categories were annotated using crowdsourcing: each image was labeled by multiple annotators with a description in the form of a verb or a verb-object pair. The main drawback of TUHOI is that 1576 out of 2974 action categories occur only once, limiting its usefulness for VSD. The Microsoft Common Objects in Context (COCO) dataset is very popular in the language/vision community, as it consists of over 120k images with extensive annotation, including labels for 91 object categories and five descriptions per image. COCO contains no explicit action annotation, but verbs and verb phrases can be extracted from the descriptions. (But note that not all the COCO images depict actions.)

The recently created Humans Interacting with Common Objects (HICO) dataset is conceptually similar to VerSe. It consists of 47774 images annotated with 111 verbs and 600 human-object interaction categories. Unlike other existing datasets, HICO uses sense-based distinctions: actions are denoted by sense-object pairs, rather than by verb-object pairs. HICO doesn't aim for complete coverage, but restricts itself to the top three WordNet senses of a verb. The dataset would be suitable for performing visual sense disambiguation, but has so far not been used in this way.

3 VerSe Dataset and Annotation

We want to build an unsupervised visual sense disambiguation system, i.e., a system that takes an image and a verb and returns the correct sense of the verb. As discussed in Section 2.1, most existing datasets are not suitable for this task, as they do not include word sense annotation. We therefore develop our own dataset with gold-standard sense annotation. The Verb Sense (VerSe) dataset is based on COCO and TUHOI and covers 90 verbs and around 3500 images. VerSe serves two main purposes: (1) to show the feasibility of annotating images with verb senses (rather than verbs or actions); (2) to function as test bed for evaluating automatic visual sense disambiguation methods.

Verb: touch

- ☐ make physical contact with, possibly with the effect of physically manipulating. They touched their fingertips together and smiled
- ☐ affect someone emotionally The president's speech touched a chord with voters.
- ☐ be or come in contact without control They sat so close that their arms touched.
- ☐ make reference to, involve oneself with They had wide-ranging discussions that touched on the situation in the Balkans.
- ☐ Achieve a value or quality Nothing can touch cotton for durability.
- ☐ Tinge; repair or improve the appearance of He touched on the paintings, trying to get the colors right.

Figure 3: Example item for depictability and sense annotation: synset definitions and examples (in blue) for the verb *touch*.

Verb Selection Action recognition datasets often use a limited number of verbs (see Table 1). We addressed this issue by using images that come with descriptions, which in the case of action images typically contain verbs. The COCO dataset includes images in the form of sentences, the TUHOI dataset is annotated with verbs or prepositional verb phrases for a given object (e.g., *sit on* chair), which we use in lieu of descriptions. We extracted all verbs from all the descriptions in the two datasets and then selected those verbs that have more than one sense in the OntoNotes dictionary, which resulted in 148 verbs in total (94 from COCO and 133 from TUHOI).

Depictability Annotation A verb can have multiple senses, but not all of them may be depictable, e.g., senses describing cognitive and perception processes. Consider two senses of *touch*: *make physical contact* is depictable, whereas *affect emotionally* describes a cognitive process and is not depictable. We therefore need to annotate the synsets of a verb as depictable or non-depictable. Amazon Mechanical Turk (AMT) workers were presented with the definitions of all the synsets of a verb, along with ex-

Verb type	Examples	Verbs	Images	Senses	Depct	ITA
Motion	run, walk, jump, etc.	39	1812	10.76	5.79	0.680
Non-motion	sit, stand, lay, etc.	51	1698	8.27	4.86	0.636

Table 2: Overview of VerSe dataset divided into motion and non-motion verbs; Depct: depictable senses; ITA: inter-annotator agreement.

amples, as given by OntoNotes. An example for this annotation is shown in Figure 3. We used OntoNotes instead of WordNet, as WordNet senses are very fine-grained and potentially make depictability and sense annotation (see below) harder. Granularity issues with WordNet for text-based WSD are well documented (Navigli, 2009).

OntoNotes lists a total of 921 senses for our 148 target verbs. For each synset, three AMT workers selected all depictable senses. The majority label was used as the gold standard for subsequent experiments. This resulted in a 504 depictable senses. Inter-annotator agreement (ITA) as measured by Fleiss' Kappa was 0.645.

Sense Annotation We then annotated a subset of the images in COCO and TUHOI with verb senses. For every image we assigned the verb that occurs most frequently in the descriptions for that image (for TUHOI, the descriptions are verb-object pairs, see above). However, many verbs are represented by only a few images, while a few verbs are represented by a large number of images. The datasets therefore show a Zipfian distribution of linguistic units, which is expected and has been observed previously for COCO (Ronchi and Perona, 2015). For sense annotation, we selected only verbs for which either COCO or TUHOI contained five or more images, resulting in a set of 90 verbs (out of the total 148). All images for these verbs were included, giving us a dataset of 3518 images: 2340 images for 82 verbs from COCO and 1188 images for 61 verbs from TUHOI (some verbs occur in both datasets).

These image-verb pairs formed the basis for sense annotation. AMT workers were presented with the image and all the depictable OntoNotes senses of the associated verb. The workers had to chose the sense of the verb that was instantiated in the image (or "none of the above", in the case of irrelevant images). Annotators were given sense definitions and examples, as for the depictability annotation (see Figure 3). For every image-verb pair, five annotators performed the sense annotation task. A total of 157 annotators participated, reaching an inter-annotator agreement of 0.659 (Fleiss' Kappa). Out of 3528 images, we discarded 18 images annotated with "none of the above", resulting in a set of 3510 images covering 90 verbs and 163 senses. We present statistics of our dataset in Table 2; we group the verbs into motion verbs and non-motion verb using Levin (1993) classes.

4 Visual Sense Disambiguation

For our disambiguation task, we assume we have a set of images I, and a set of polysemous verbs V and each image $i \in I$ is paired with a verb $v \in V$. For example, Figure 1 shows different images paired with the verb *play*. Every verb $v \in V$, has a set of senses $\mathcal{S}(v)$, described in a dictionary $\mathcal{D}$. Now given an image i paired with a verb v, our task is to predict the correct sense $\hat{s} \in \mathcal{S}(v)$, i.e., the sense that is depicted by the associated image. Formulated as a scoring task, disambiguation consists of finding the maximum over a suitable scoring function Φ:

$$\hat{s} = \arg\max_{s \in \mathcal{S}(v)} \Phi(s, i, v, \mathcal{D}) \quad (1)$$

For example, in Figure 1, the correct sense for the first image is *participate in sport*, for the second one it is *play on an instrument*, etc.

The Lesk (1986) algorithm is a well known knowledge-based approach to WSD which relies on the calculation of the word overlap between the sense definition and the context in which a word occurs. It is therefore an unsupervised approach, i.e., it does not require sense-annotated training data, but instead exploits resources such as dictionaries or ontologies to infer the sense of a word in context. Lesk uses the following scoring function to disambiguate the sense of a verb v:

$$\Phi(s, v, \mathcal{D}) = |\mathit{context}(v) \cap \mathit{definition}(s, \mathcal{D})| \quad (2)$$

Here, $\mathit{context}(v)$ the set of words that occur close the target word v and $\mathit{definition}(s, \mathcal{D})$ is the set of words in the definition of sense s in the dictionary $\mathcal{D}$. Lesk's approach is very sensitive to the exact wording of definitions and results are known to change dramatically for different sets of definitions (Navigli, 2009). Also, sense definitions are often very

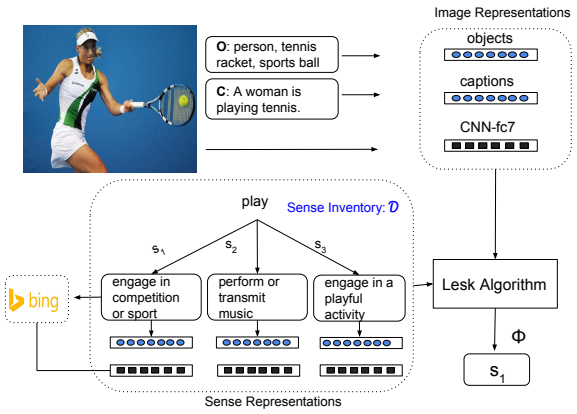

Figure 4: Schematic overview of the visual sense disambiguation model.

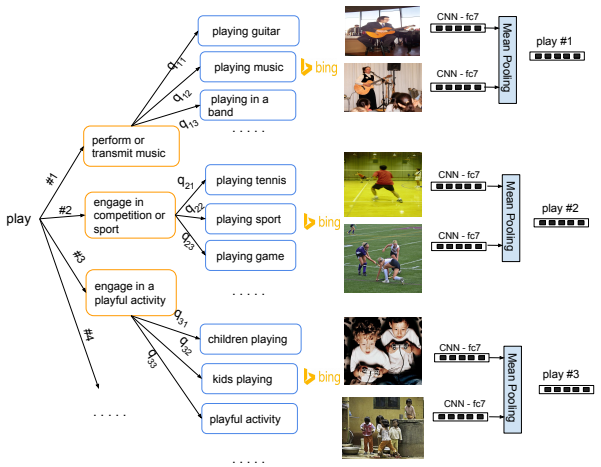

Figure 5: Extracting visual sense representation for the verb *play*.

short and do not provide sufficient vocabulary or context.

We propose a new variant of the Lesk algorithm to disambiguate the verb sense that is depicted in an image. In particular, we explore the effectiveness of textual, visual and multimodal representations in conjunction with Lesk. An overview of our methodology is given in Figure 4. For a given image i labeled with verb v (here *play*), we create a representation (the vector $\mathbf{i}$), which can be text-based (using the object labels and descriptions for i), visual, or multimodal. Similarly, we create text-based, visual, and multimodal representations (the vector $\mathbf{s}$) for every sense s of a verb. Based on the representations $\mathbf{i}$ and $\mathbf{s}$ (detailed below), we can then score senses as:[1]

$$\Phi(s, v, i, \mathcal{D}) = \mathbf{i} \cdot \mathbf{s} \quad (3)$$

Note that this approach is unsupervised: it requires no sense annotated training data; we will use the sense annotations in our VerSe dataset only for evaluation.

4.1 Sense Representations

For each candidate verb sense, we create a text-based sense representation $\mathbf{s^t}$ and a visual sense representation $\mathbf{s^c}$.

Text-based Sense Representation We create a vector $\mathbf{s^t}$ for every sense $s \in \mathcal{S}(v)$ of a verb v from its definition and the example usages provided in the OntoNotes dictionary $\mathcal{D}$. We apply word2vec (Mikolov et al., 2013), a widely used model of word embeddings, to obtain a vector for every content word in the definition and examples of the sense. We then take the average of these vectors to compute an overall representation of the verb sense. For our experiments we used the pre-trained 300 dimensional vectors available with the word2vec package (trained on part of Google News dataset, about 100 billion words).

Visual Sense Representation Sense dictionaries typically provide sense definitions and example sentences, but no visual examples or images. For nouns, this is remedied by ImageNet (Deng et al., 2009), which provides a large number of example images for a subset of the senses in the WordNet noun hierarchy. However, no comparable resource is available for verbs (see Section 2.1).

In order to obtain visual sense representation $\mathbf{s^c}$, we therefore collected sense-specific images for the verbs in our dataset. For each verb sense s, three trained annotators were presented with the definition and examples from OntoNotes, and had to formulate a query $Q(s)$ that would retrieve images depicting the verb sense when submitted to a search engine. For every query q we retrieved images $I(q)$ using Bing image search (for examples, see Figure 5). We used the top 50 images returned by Bing for every query.

Once we have images for every sense, we can turn these images into feature representations us-

[1] Taking the dot product of two normalized vectors is equivalent to using cosine as similarity measure. We experimented with other similarity measures, but cosine performed best.

ing a convolutional neural network (CNN). Specifically, we used the VGG 16-layer architecture (VGGNet) trained on 1.2M images of the 1000 class ILSVRC 2012 object classification dataset, a subset of ImageNet (Simonyan and Zisserman, 2014). This CNN model has a top-5 classification error of 7.4% on ILSVRC 2012. We use the publicly available reference model implemented using CAFFE (Jia et al., 2014) to extract the output of the fc7 layer, i.e., a 4096 dimensional vector $\mathbf{c_i}$, for every image i. We perform mean pooling over all the images extracted using all the queries of a sense to generate a single visual sense representation $\mathbf{s^c}$ (shown in Equation 4):

$$\mathbf{s^c} = \frac{1}{n} \sum_{q_j \in Q(s)} \sum_{i \in I(q_j)} \mathbf{c_i} \quad (4)$$

where n is the total number of images retrieved per sense s.

4.2 Image Representations

We first explore the possibility of representing the image indirectly, viz., through text associated with it in the form of object labels or image descriptions (as shown in Figure 4). We experiment with two different forms of textual annotation: GOLD annotation, where object labels and descriptions are provided by human annotators, and predicted (PRED) annotation, where state-of-the-art object recognition and image description generation systems are applied to the image.

Object Labels (O) GOLD object annotations are provided with the two datasets we use. Each image sampled from COCO is annotated with one or more of 91 object categories. Each image from TUHOI is annotated with one more of 189 object categories. PRED object annotations were generated using the same VGG-16-layer CNN object recognition model that was used to compute visual sense representations. Only object labels with object detection threshold of $t > 0.2$ were used.

Descriptions (C) To obtain GOLD image descriptions, we used the used human-generated descriptions that come with COCO. For TUHOI images, we generated descriptions of the form subject-verb-object, where the subject is always *person*, and the verb-object pairs are the action labels that come with TUHOI. To obtain PRED descriptions, we generated three descriptions for every image using the state-of-the-art image description system of Vinyals et al. (2015).[2]

We can now create a textual representation $\mathbf{i^t}$ of the image i. Again, we used word2vec to obtain word embeddings, but applied these to the object labels and to the words in the image descriptions. An overall representation of the image is then computed by averaging these vectors over all labels, all content words in the description, or both.

Creating a visual representation $\mathbf{i^c}$ of an image i is straightforward: we extract the fc7 layer of the VGG-16 network when applied to the image and use the resulting vector as our image representation (same setup as in Section 4.1).

Apart from experimenting with separate textual and visual representations of images, it also makes sense to combine the two modalities into a multimodal representation. The simplest approach is a concatenation model which appends textual and visual features. More complex multimodal vectors can be created using methods such as Canonical Correlation Analysis (CCA) and Deep Canonical Correlation Analysis (DCCA) (Hardoon et al., 2004; Andrew et al., 2013; Wang et al., 2015). CCA allows us to find a latent space in which the linear projections of text and image vectors are maximally correlated (Gong et al., 2014; Hodosh et al., 2015). DCCA can be seen as non-linear version of CCA and has been successfully applied to image description task (Yan and Mikolajczyk, 2015), outperforming previous approaches, including kernel-based CCA.

We use both CCA and DCCA to map the vectors $\mathbf{i^t}$ and $\mathbf{i^c}$ (which have different dimensions) into a joint latent space of n dimensions. We represent the projected vectors of textual and visual features for image i as $\mathbf{i^{t'}}$ and $\mathbf{i^{c'}}$ and combine them to obtain multimodal representation $\mathbf{i^m}$ as follows:

$$\mathbf{i^m} = \lambda_t \mathbf{i^{t'}} + \lambda_c \mathbf{i^{c'}} \quad (5)$$

We experimented with a number of parameter settings for λ_t and λ_c for textual and visual models respectively. We use the same model to combine the multimodal representation for sense s as follows:

$$\mathbf{s^m} = \lambda_t \mathbf{s^{t'}} + \lambda_c \mathbf{s^{c'}} \quad (6)$$

[2]We used Karpathy's implementation, publicly available at `https://github.com/karpathy/neuraltalk`.

We use these vectors ($\mathbf{i^t}$, $\mathbf{s^t}$), ($\mathbf{i^c}$, $\mathbf{s^c}$) and ($\mathbf{i^m}$, $\mathbf{s^m}$) as described in Equation 3 to perform sense disambiguation.

5 Experiments

5.1 Unsupervised Setup

To train the CCA and DCCA models, we use the text representations learned from image descriptions of COCO and Flickr30k dataset as one view and the VGG-16 features from the respective images as the second view. We divide the data into train, test and development samples (using a 80/10/10 split). We observed that the correlation scores for DCCA model were better than for the CCA model. We use the trained models to generate the projected representations of text and visual features for the images in VerSe. Once the textual and visual features are projected, we then merge them to get the multimodal representation. We experimented with different ways of combining visual and textual features projected using CCA or DCCA: (1) weighted interpolation of textual and visual features (see Equations 5 and 6), and (2) concatenating the vectors of textual and visual features.

To evaluate our proposed method, we compare against the first sense heuristic, which defaults to the sense listed first in the dictionary (where senses are typically ordered by frequency). This is a strong baseline which is known to outperform more complex models in traditional text-based WSD. In VerSe we observe skewness in the distribution of the senses and the first sense heuristic is as strong as over text. Also the most frequent sense heuristic, which assigns the most frequently annotated sense for a given verb in VerSe, shows very strong performance. It is supervised (as it requires sense annotated data to obtain the frequencies), so it should be regarded as an upper limit on the performance of the unsupervised methods we propose (also, in text-based WSD, the most frequent sense heuristic is considered an upper limit, Navigli (2009)).

5.1.1 Results

In Table 3, we summarize the results of the gold-standard (GOLD) and predicted (PRED) settings for motion and non-motion verbs across representations. In the GOLD setting we find that for both types of verbs, textual representations based on image descriptions (C) outperform visual representations (CNN features). The text-based results compare favorably to the original Lesk (as described in Equation 2), which performs at 30.7 for motion verbs and 36.2 for non-motion verbs in the GOLD setting. This improvement is clearly due to the use of word2vec embeddings.[3] Note that CNN-based visual features alone performed better than gold-standard object labels alone in the case of motion verbs.

We also observed that adding visual features to textual features improves performance in some cases: multimodal features perform better than textual features alone both for object labels (CNN+O) and for image descriptions (CNN+C). However, adding CNN features to textual features based on object labels and descriptions together (CNN+O+C) resulted in a small decrease in performance. Furthermore, we note that CCA models outperform simple vector concatenation in case of GOLD setting for motion verbs, and overall DCCA performed considerably worse than concatenation. Note that for CCA and DCCA we report the best performing scores achieved using weighted interpolation of textual and visual features with weights $\lambda_t = 0.5$ and $\lambda_c = 0.5$.

When comparing to our baseline and upper limit, we find that the all the GOLD models which use descriptions-based representations (except DCCA) outperform to the first sense heuristic for motion-verbs (accuracy 70.8), whereas they performed below the first sense heuristic in case of non-motion verbs (accuracy 80.6). As expected, both motion and non-motion verbs performed significantly below the most frequent sense heuristic (accuracy 86.2 and 90.7 respectively), which we argued provides an upper limit for unsupervised approaches.

We now turn the PRED configuration, i.e., to results obtained using object labels and image descriptions predicted by state-of-the-art automatic systems. This is arguably the more realistic scenario, as it only requires images as input, rather than assuming human-generated object labels and image descriptions (though object detection and image description systems are required instead). In the PRED setting, we find that textual features based on ob-

[3] We also experimented with Glove vectors (Pennington et al., 2014) but observed that word2vec representations consistently achieved better results that Glove vectors.

(a) Motion verbs (39), FS: 70.8, MFS: 86.2

Annotation	Textual			Vis	Concat (CNN+)			CCA (CNN+)			DCCA (CNN+)		
	O	C	O+C	CNN	O	C	O+C	O	C	O+C	O	C	O+C
GOLD	54.6	**73.3**	**75.6**	58.3	66.6	**74.7**	**73.8**	50.5	**75.4**	**74.0**	52.4	66.3	68.3
PRED	65.1	54.9	61.6	58.3	**72.6**	63.6	66.5	54.0	56.6	56.2	57.1	56.5	56.2

(b) Non-motion verbs (51), FS: 80.6, MFS: 90.7

Annotation	Textual			Vis	Concat (CNN+)			CCA (CNN+)			DCCA (CNN+)		
	O	C	O+C	CNN	O	C	O+C	O	C	O+C	O	C	O+C
GOLD	57.0	72.7	72.6	56.1	66.0	72.2	71.3	53.6	71.6	70.2	57.3	59.8	55.1
PRED	59.0	64.3	64.0	56.1	63.8	66.3	66.1	50.7	55.3	54.8	49.5	50.0	50.0

Table 3: Accuracy scores for motion and non-motion verbs using for different types of sense and image representations (O: object labels, C: image descriptions, CNN: image features, FS: first sense heuristic, MFS: most frequent sense heuristic). Configurations that performed better than FS in **bold**.

Motion verbs (19), FS: 60.0, MFS: 76.1				
Features	GOLD		PRED	
	Sup	Unsup	Sup	Unsup
O	82.3	35.3	80.0	43.8
C	78.4	53.8	69.2	41.5
O+C	80.0	55.3	70.7	45.3
CNN	82.3	58.4	82.3	58.4
CNN+O	83.0	48.4	83.0	60.0
CNN+C	82.3	66.9	82.3	53.0
CNN+O+C	83.0	58.4	83.0	55.3

Table 4: Accuracy scores for motion verbs for both supervised and unsupervised approaches using different types of sense and image representation features.

Non-Motion verbs (19), FS: 71.3, MFS: 80.0				
Features	GOLD		PRED	
	Sup	Unsup	Sup	Unsup
O	79.1	48.6	78.2	46.0
C	79.1	53.9	77.3	61.7
O+C	79.1	66.0	77.3	55.6
CNN	80.0	55.6	80.0	55.6
CNN+O	80.0	56.5	80.0	52.1
CNN+C	80.0	56.5	80.3	60.0
CNN+O+C	80.0	59.1	80.0	55.6

Table 5: Accuracy scores for non-motion verbs for both supervised and unsupervised approaches using different types of sense and image representation features.

ject labels (O) outperform both first sense heuristic and textual features based on image descriptions (C) in the case of motion verbs. Combining textual and visual features via concatenation improves performance for both motion and non-motion verbs. The overall best performance of 72.6 for predicted features is obtained by combining CNN features and embeddings based on object labels and outperforms first sense heuristic in case of motion verbs (accuracy 70.8). In the PRED setting for both classes of verbs the simpler concatenation model performed better than the more complex CCA and DCCA models. Note that for CCA and DCCA we report the best performing scores achieved using weighted interpolation of textual and visual features with weights $\lambda_t = 0.3$ and $\lambda_c = 0.7$. Overall, our findings are consistent with the intuition that motion verbs are easier to disambiguate than non-motion verbs, as they are more depictable and more likely to involve objects. Note that this is also reflected in the higher inter-annotator agreement for motion verbs (see Table 2).

5.2 Supervised Experiments and Results

Along with the unsupervised experiments we investigated the performance of textual and visual representations of images in a simplest supervised setting. We trained logistic regression classifiers for sense prediction by dividing the images in VerSe dataset into train and test splits. To train the classifiers we selected all the verbs which has atleast 20 images annotated and has at least two senses in VerSe. This resulted in 19 motion verbs and 19 non-motion verbs. Similar to our unsupervised experiments we explore multimodal features by using both textual and visual features for classification (similar to concatenation in unsupervised experiments).

Verb	Image	Predicted Descriptions	Pred. Obj.
play		A man holding a nintendo wii game controller. A man and a woman playing a video game. A man and a woman are playing a video game.	person, bassoon, violin fiddle, oboe, hautboy
swing		A woman standing next to a fire hydrant. A woman walking down a street holding an umbrella. A woman standing on a sidewalk holding an umbrella.	person, horizontal bar, high bar, pole
feed		A couple of cows standing next to each other. A cow that is standing in the dirt. A close up of a horse in a stable	arabian camel, dromedary, person

Table 6: Images that were assigned an incorrect sense in the PRED setting.

In Table 4 we report accuracy scores for 19 motion verbs using a supervised logistic regression classifier and for comparison we also report the scores of our proposed unsupervised algorithm for both GOLD and PRED setting. Similarly in Table 5 we report the accuracy scores for 19 non-motion verbs. We observe that all supervised classifiers for both motion and non-motion verbs performing better than first sense baseline. Similar to our findings using an unsupervised approach we find that in most cases multimodal features obtained using concatenating textual and visual features has outperformed textual or visual features alone especially in the PRED setting which is arguably the more realistic scenario. We observe that the features from PRED image descriptions showed better results for non-motion verbs for both supervised and unsupervised approaches whereas PRED object features showed better results for motion verbs. We also observe that supervised classifiers outperform most frequent sense for motion verbs and for non-motion verbs our scores match with most frequent sense heuristic.

5.3 Error Analysis

In order to understand the cases where the proposed unsupervised algorithm failed, we analyzed the images that were disambiguated incorrectly. For the PRED setting, we observed that using predicted image descriptions yielded lower scores compared to predicted object labels. The main reason for this is that the image description system often generates irrelevant descriptions or descriptions not related to the action depicted, whereas the object labels predicted by the CNN model tend to be relevant. This highlights that current image description systems still have clear limitations, despite the high evaluation scores reported in the literature (Vinyals et al., 2015; Fang et al., 2015). Examples are shown in Table 6: in all cases human generated descriptions and object labels that are relevant for disambiguation, which explains the higher scores in the GOLD setting.

6 Conclusion

We have introduced the new task of visual verb sense disambiguation: given an image and a verb, identify the verb sense depicted in the image. We developed the new VerSe dataset for this task, based on the existing COCO and TUHOI datasets. We proposed an unsupervised visual sense disambiguation model based on the Lesk algorithm and demonstrated that both textual and visual information associated with an image can contribute to sense disambiguation. In an in-depth analysis of various image representations we showed that object labels and visual features extracted using state-of-the-art convolutional neural networks result in good disambiguation performance, while automatically generated image descriptions are less useful.

References

Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1247–1255.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433.

Kobus Barnard, Matthew Johnson, and David Forsyth. 2003. Word sense disambiguation with pictures. In *Proceedings of the HLT-NAACL 2003 workshop on Learning word meaning from non-linguistic data-Volume 6*, pages 1–5. Association for Computational Linguistics.

Samuel Brody and Mirella Lapata. 2008. Good neighbors make good senses: Exploiting distributional similarity for unsupervised wsd. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 65–72. Association for Computational Linguistics.

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. HICO: A benchmark for recognizing human-object interactions in images. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1017–1025.

Xinlei Chen, Alan Ritter, Abhinav Gupta, and Tom M. Mitchell. 2015. Sense discovery via co-clustering on images and text. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5298–5306.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2015. The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.

Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1473–1482.

Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, pages 529–545.

David R. Hardoon, Sándor Szedmák, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2015. Framing image description as a ranking task: Data, models and evaluation metrics (extended abstract). In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 4188–4192.

Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2006. Ontonotes: The 90% solution. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA*, pages 57–60.

Sujay Kumar Jauhar, Chris Dyer, and Eduard H. Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 683–693.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, pages 675–678.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137.

Adam Kilgarrif. 1998. Senseval: An exercise in evaluating word sense disambiguation programs. In *Proc. of the first international conference on language resources and evaluation*, pages 581–588.

Dieu Thu Le, Raffaella Bernardi, and Jasper Uijlings. 2013. Exploiting language models to recognize unseen actions. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 231–238. ACM.

Dieu-Thu Le, Jasper Uijlings, and Raffaella Bernardi, 2014. *Proceedings of the Third Workshop on Vision and Language*, chapter TUHOI: Trento Universal Human Object Interaction Dataset, pages 17–24. Dublin City University and the Association for Computational Linguistics.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC 1986, Toronto, Ontario, Canada, 1986*, pages 24–26.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 64–71. Association for Computational Linguistics.

Nicolas Loeff, Cecilia Ovesdotter Alm, and David A. Forsyth. 2006. Discriminating image senses by clustering with multimodal features. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, pages 547–554. Association for Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 279–286. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Matteo Ruggero Ronchi and Pietro Perona. 2015. Describing common human visual actions in images. In *Proceedings of the British Machine Vision Conference (BMVC 2015)*, pages 52.1–52.12. BMVA Press, September.

Sascha Rothe and Hinrich Schutze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1793–1803.

Kate Saenko and Trevor Darrell. 2008. Unsupervised learning of visual sense models for polysemous words. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 1393–1400.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff A. Bilmes. 2015. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1083–1092.

Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3441–3450.

Bangpeng Yao and Li Fei-Fei. 2010. Grouplet: A structured image representation for recognizing human and object interactions. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 9–16. IEEE.

Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. 2011. Human action recognition by learning bases of action attributes and parts. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1331–1338. IEEE.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, System Demonstrations*, pages 78–83.